Balanced-Delay Filterbank for Closed-Loop Spatial Audio Coding

Ikhwana Elfitri, Heru Dibyo Laksono, and Al Kautsar Permana
Department of Electrical Engineering, Faculty of Engineering, Andalas University,
Kampus UNAND Limau Manih, Padang, 25163, Sumatera Barat, Indonesia
E-mail: ikhwana@ft.unand.ac.id

***Abstract*—Closed-loop configuration has been introduced to Spatial Audio Coding (SAC) which make it possible to minimise distortion introduced during the quantisation and encoding processes. Significant performance improvement has been shown and reported in some papers. However, implementation of closed-loop system directly to MPEG Surround (MPS) is still problematic due to the unbalanced-delay filterbank that is used in the MPEG standard which is not appropriate for a closed-loop system that needs synchronisation between the original audio signals with the target audio signals. In this paper, we investigate the delay characteristic of the Quadrature Mirror Filterbank (QMF) which is used in several MPEG standards. Based on this study, a balanced-delay QMF is proposed and tested in a closed-loop MPS system. The results of the experiments show that approximately 8 dB of SNR improvement is achieved when applying balanced-delay filterbank compared to the unbalanced-delay filterbank as specified in MPS.**

***Keywords*—Filterbank, Spatial Audio Coding, Multichannel Audio Signals, Closed-Loop System**

I. Introduction

Spatial audio [1] have been considered, since at least the last decade, as the future of high quality audio reproduction where much more realistic sound scene can be created by means of an array of a number loudspeakers. Even three-dimensional (3D) audio rendering is possible with the introduction of higher number of applicable loudspeakers such as the 22.2 audio system [2], [3]. The technology of spatial audio will be complement to the advancement of high definition and 3D video technology as high quality audiovisual and multimedia technology. For the purpose of processing and encoding, spatial audio is normally represented as multichannel audio signals [4]–[6].

In response to the increasing development of spatial audio, Moving Picture Expert Group (MPEG) has already released two standards that are aimed as compression tools for multichannel audio signals. The first one is MPEG Surround (MPS) [7]–[10] that works based on the principle of Spatial Audio Coding (SAC) [11]–[14] in which multichannel audio signals can be downmixed into a mono or stereo audio signals and then transmitted with spatial parameters as side information. The second standard is MPEG Spatial Audio Object Coding (SAOC) [15]–[18] which has capability of providing user interaction to remix the composition of the rendered spatial audio scene. With this SAOC standard, it is possible to render music composition as vocal only or background music only. In addition, MPEG is currently finalising its standard for encoding 3D audio [19] which combined both channel-based and object-based audio reproduction. However, both MPS and MPEG SAOC audio codec standards seem to be powerful only at low bit rate applications although the availability of high speed network connection has motivated to apply much more high quality spatial audio applications for accompanying high definition TV and other video applications.

In an attempt to improve the performance of existing MPS standard, Closed-Loop Spatial Audio Coding (CL-SAC) [20]–[22] has been proposed with an ability to minimise distortion due to quantisation and encoding processes. Some benefits of CL-SAC can be considered. First, significant performance improvement has been reported while maintaining all features that are belong to MPS. Second, it is bitrate scalable up to a bitrate that can achieve perfect waveform reconstruction. Third, all features of MPS that works based on the principle of spatial audio coding such as backward compatibility and binarual rendering, can be maintained. However, the CL-SAC has been employed in the frequency domain by transforming audio signal using Modified Discrete Cosine Transform (MDCT), thus it is not applicable directly to MPS. In addition, processing audio signal in the frequency domain possibly introduce more artifact due to having lower time resolution.

In this paper we discuss the delay characteristic of QMF filterbank that used in MPS as a reason why the CL-SAC cannot be implemented in the domain of QMF filterbank as MPS does. Then, we propose a balanced-delay filterbank that can solve the problem encountered in applying closed-loop system to MPS. In the next section an overview of closed-loop spatial audio coding and unbalanced-delay filterbank in MPS standard will be discussed, followed by discussion on its delay characteristic and a balanced-delay solution proposed in this work. In the end of the paper, the results of the experiments are presented which is finalised with the conclusion.

II. Overview of Closed-Loop Framework for Spatial Audio Coding

Spatial Audio Coding (SAC) is an advance encoding technique having advantage of efficient compression of multichannel audio signals. The encoding process can be considered as: first, extracting basic parameters which are commonly known as spatial parameters, and second, reducing the number of audio channels into a single (mono) or double (stereo) audio signals which is usually referred to as downmixing process. The benefit of encoding using this method is that the downmixed signals can be compressed by any audio encoding technique while the spatial parameters can be transmitted conditionally depending on various circumstances such as the

978-1-4799-7711-6/15/$31.00 © 2015 IEEE

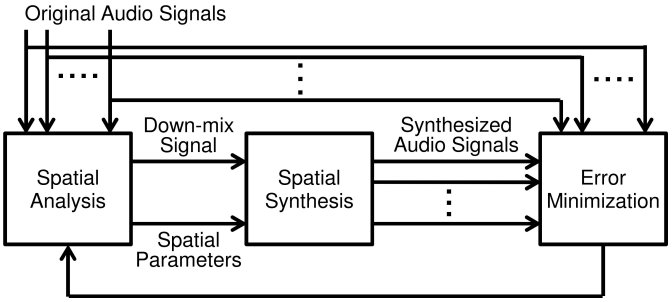

Fig. 1. Framework of Closed-Loop Spatial Audio Coding (CL-SAC).

desired operating bit rate. Moreover, SAC also benefits of being able to put spatial parameters as the side information thus it can be removed if needed while it is still able to decode only the downmix signals, hence, it is favourable for upgrading existing audio system which has been widely implemented to have multichannel extension. This is known as a backward compatibility. In the decoding process, the downmix signals can be extended back into channel configurations that can be either identical or different from the one fed in the encoder side.

With a main purpose of improving its performance, SAC technique has been configured as a closed-loop scheme. A framework of the Closed-Loop SAC (CL-SAC) encoding process, shown in Fig. 1, can be discussed as follows. In contrast to a single spatial analysis block that is normally employed in an open-loop encoding system, three blocks: spatial analysis, spatial synthesis, and error minimisation, are applied in a closed-loop encoding approach. The spatial synthesis block, that must be similar to the one applied in the decoder side, is required as a model so the whole encoder system is capable of theoritically reconstructing multichannel audio signals. Thus, an error minimisation block can be very useful to compare the input audio signals with the reconstructed signals. Based on this comparison, some sort of minimisation can be employed with suitable error criteria. However, in order to make a fair comparison between the original and the reconstructed audio signals, both of them have to be synchronised. For this reason due to the unbalanced time delay characteristic of filterbank specified in MPS standard, the CL-SAC is applied in the frequency domain. Even though increasing performance have been shown [20]–[22], however, further improving is expected when the CL-SAC can be applied in the domain of filterbank.

III. Unbalanced-Delay Filterbank Specified in MPEG Surround Standard

MPEG Surround is an international standard for encoding multichannel audio developed based on the principle of SAC. Even though having so many interesting features and functionalities, it is still possible to increase MPS performance. Employing MPS as a closed-loop configuration has shown that significant performance improvement can be achieved. However, due to unbalanced-delay introduced by both analysis and synthesis filterbank used in MPS, it is difficult to synchronise the synthesised audio signals with the original signals. Hence, the spatial analysis and synthesis are performed in the frequency domain by using MDCT transformation. In this section we discuss in details the drawback of QMF filterbank in terms of applying closed-loop configuration and then present a balanced-delay filterbank as a solution.

MPS standard specifies a hybrid Quadrature Mirror Filterbank (QMF) [23]–[25] which is aimed at representing audio signal in a perceptual time-frequency domain. In general, QMF consists of an analysis and synthesis filterbank. An Analysis QMF (A-QMF) is useful to decompose audio signal in each channel into 71 hybrid (non-uniform) subband signals, while a Synthesis QMF (S-QMF) is applied to turn back the audio signal into its time domain signal representation. The initial process in the A-QMF, as given in Fig. 2, is a convolution of audio signal, $s[i]$, with a set of filters having impulse responses, $G_{0,m_0}[i]$ [26], given as,

$$G_{0,m_0}[i] = h_0[i] exp\left[j\frac{\pi}{2M_0}(m_0+0.5)(2i-1)\right] \quad (1)$$

where $h_0[i]$ is a common core filter having a length of 640, m_0 is the index of the filter, and $M_0 = 64$ is the total number of filters. Each filter in the filterbank produces audio signal, having uniform (equal-bandwidth), called subband signal. Furthermore, each uniform subband signal is proceeded to a downsampler with a factor of M_0 to obtain lower rate signals, hence, can greatly reduce the complexity of subsequent processing of the subband signals. Thus, the signal created from the downsampling process, $x_{m_0}[n]$, can be represented as,

$$x_{m_0}[n] = (s[i] * G_{0,m_0}[i])[M_0 n] \quad (2)$$

The final stage of the A-QMF aims to further reduce the bandwidth of the subband signals at low-frequency regions while maintaining the bandwidth of the subband signals at high-frequency region as they were. For this purpose the first subband signal is convolved with a set of filter having impulse responses, $G_{1,m_1}[n]$ [26], given by,

$$G_{1,m_1}[n] = h_1[n] exp\left[j\frac{\pi}{4}(m_1+0.5)(n-6)\right] \quad (3)$$

where m_1 ($m_1 = 0, 1, ..., 7$) is the index of the filter, and $h_1[n]$ is the prototype filter with a length of 13. These concolution processes create as many as 8 sub-subband signals. However, note that at this stage, filtering processes cause coincidence of some passband with the other stopbands, hence output of the $G_{1,2}$ and $G_{1,2}$ are combined together while output of the $G_{1,3}$ and $G_{1,4}$ are also combined. As a result there only 6 sub-subbands created from the first subband.

The second and the third subband signals are also processed with another filterbank with impulse responses, $G_{2,m_2}[n]$ [26], given by,

$$G_{2,m_2}[n] = h_2[n] cos\left[\pi(m_2-1)(n-6)\right] \quad (4)$$

where $m_2, (m_2 = 0, 1)$, thus, every subband signal obtains 2 sub-subbands. The prototype filter, $h_2[n]$, also has a filter length equal to 13.

From this final stage we can see that 3 subband signals have been decomposed into 10 sub-subbands while the other 61 subband signals are not processed and kept as they were. However, the 10 sub-subband signals have been delayed due to filtering process. In order to make them all having identical delays the other 61 subband signal need to be delayed by the same time as the filter does. At this point, all 10 sub-subband

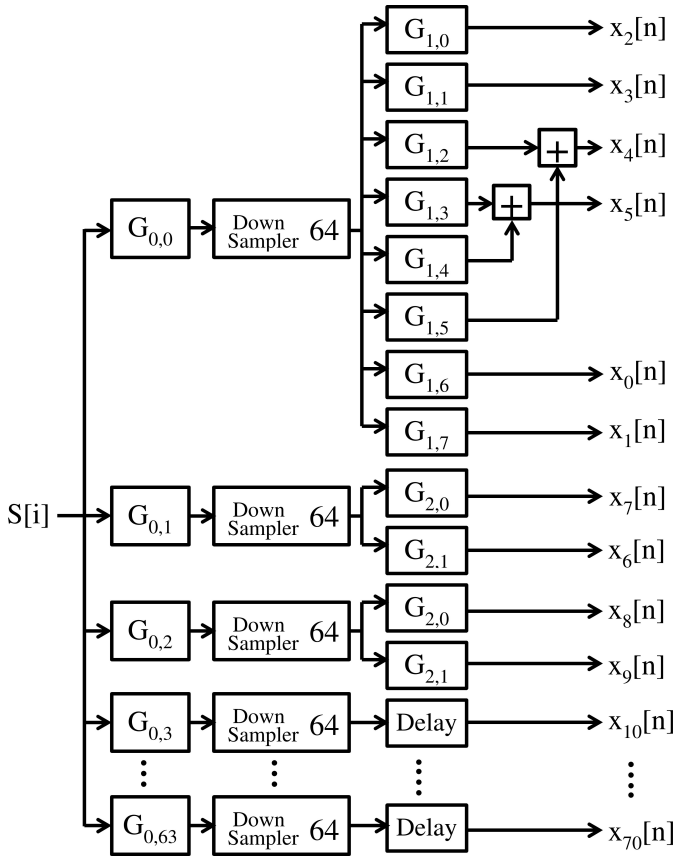

Fig. 2. Block diagram of Analysis Quadrature Mirror Filterbank (A-QMF)

and 61 subband signals are referred to as hybrid band signals forming 71 hybrid bands of A-QMF.

As specified in MPS standard, the hybrid band signals go through calculation of spatial parameters and reduction of channel numbers by downmixing process. The resulting audio signals need to be transformed back to time domain by applying the hybrid band signals to the S-QMF. As done in the A-QMF, the transformation of audio signal to time domain can be done in 2 steps. First step is summation of the first 10 hybrid band signals, which are basically sub-subband signals, to transform them back to be 3 original subband signals. Subsequently, the newly formed 3 subband signals, along with the remaining 61 subband signals, representing the original 64 subband signals, are filtered with impulse responses, $H_{m_0}[i]$ [26], as follows:

$$H_{0,m_0}[i] = \frac{1}{M_0} h_0[i] exp\left[j\frac{\pi}{2M_0}(m_0 + 0.5)(2i - 255) \right] \quad (5)$$

where $h_0[i], m_0$, and M_0 are defined similar to those in (1).

IV. The Proposed Balanced-Delay Filterbank

The main problem in applying CL-SAC in the domain of hybrid filterbank is the unbalanced-delay introduced by the analysis and synthesis filterbank. The A-QMF delays audio signals as many as 704 samples in the time domain (see impulse responses of the analysis filterbank, $G_{0,s}$) which is equivalent to 11 samples in the QMF subband domain. In contrast, the S-QMF provides audio signals that are delayed as many as 257 samples in the time domain (refer to impulse responses of the analysis filterbank, $H_{0,s}$) which is identical to fractional 4.015625 samples in the QMF subband domain [26]. For this reason, comparing the original audio signals to the reconstructed audio signals cannot be much useful in the CL-SAC.

TABLE I. SNRs (in decibel) of Various Signal Transformation Schemes Using QMF Filterbank

Scheme No	**(1)**	**(2)**	**(3)**	**(4)**
Audio Excerpt	**Subband-Tx**	**Time-Tx**	**Cont-SAC**	**OL-SAC**
Acoustical music	61.36	55.33	44.76	23.91
Clapping hands	59.42	53.41	44.96	26.25
Classical music	58.97	52.96	36.00	21.86
People laughing	59.49	53.48	40.36	24.74
Phone talking	59.77	53.44	43.04	24.75

The disadvantages of the QMF filterbank due to the unbalanced-delay properties can be discussed referring to Table I where Signal-to-Noise Ratio (SNR) from a number signal transformation schemes using QMF for 5 audio signals are presented. In the first scheme the audio signals are transformed from time to subband domain on the encoder side before being transmitted to the decoder. On the decoder side the subband signals are transformed back to the time domain. This scheme is denoted as Subband-Tx. The second scheme, named as Time-Tx, is performed by transforming the audio signals from time to subband domain on the encoder side. Furthermore, without performing the spatial analysis block the audio signals are transformed back to the time domain and then transmitted to the decoder. The reverse process is performed on the decoder side. The third scheme is called Cont-SAC. It uses the MPS system without performing any coding or quantisation process. This means that all signals and parameters are transmitted in their original form. The fourth, called OL-SAC, is a scheme that applies the MPS system. The spatial parameters are quantised and then transmitted to the decoder. However, the original form of the downmix and residual signals are transmitted instead of the encoded form.

It can be seen in Table I that there is a trend of SNR degradation, for each audio excerpt, when using the Time-Tx compared to the Subband-Tx scheme. It suggests that both the transformation of the audio signals from subband domain to time domain on the encoder side and the transformation of the audio signals from time domain to subband domain on the decoder side introduce signal error. Furthermore, the Table also shows that the OL-SAC has much lower SNRs than the Cont-SAC. It suggests that the OL-SAC applied in the QMF domain considerably suffers from distortion which possibly not only introduced by the quantisation of the spatial parameters but also, as previously discussed, caused by different lengths of delay introduced by the analysis and synthesis filterbanks.

The delay introduced by S-QMF can be discussed in details as follows. The standard filterbank can be simply modified to make it introduce different time delay. However, the losses caused by filtering the audio signal can be higher. In Fig. 3, it is shown that the SNR of the filterbank, in this case the SNR was taken when an audio signal of Classical music was used as input, varies when we try to modify the delay introduced by the S-QMF. It is understandable why MPS is specified to have S-QMF with time delay of 257 samples because it provides the

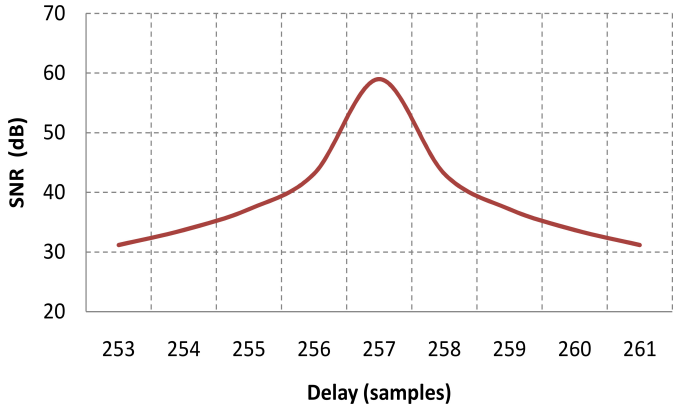

Fig. 3. SNR of Synthesis Quadrature Mirror Filterbank (S-QMF) against delay (in samples).

highest SNR even though it introduces fractional time delay. For the other scenarios when the time delay is more or less than 257 samples, we can see that the SNR decreases.

The balanced-delay filterbank in this work was simply created by replacing the term $(2i - 255)$ with $(2i - 253)$ in the standard filterbank in (5) to introduce a time delay equal to 256 samples in the time domain so that the delay in the subband domain is no longer fractional. Although SNR of the balanced-delay filterbank in an open-loop configuration can be lower than the standard filterbank, however, when employing in the closed-loop configuration its performance is expected to be better than the standard filterbank in the open-loop system.

V. Performance Evaluation

A. Experimental Setup

Five different 5-channel audio signals with duration of 12 s, consisting of: phone talking (speeches), acoustical and classical music, as well as sounds of people clapping hands and laughing, were used in the experiments. The open-loop and closed-loop structures were tested on two filterbanks: unbalanced-delay (standard) filterbank as specified in MPS standard and balanced-delay filterbank. In the experiment, the CLD and ICC are quantised as 5 and 3 bits, respectively, while the down-mix and residual signals are transmitted in continuous form without encoding. Three scenarios were tested: 2, 4 and 5 audio input channels. In order to keep consistent SNR measurement from different scenario, audio signals from the left front and left surround channels were measured.

B. SNR of Closed-Loop Spatial Audio Coding Using the Proposed Balanced-Delay Filterbank

The results are given in Fig. 4 showing that, first, as expected the open-loop MPS using balanced-delay filterbank tends to provide lower SNR than the open-loop MPS using unbalanced-delay filterbank. However, the differences of the SNRs are not significant. Second, the closed-loop structure using balanced-delay filterbank achieves higher SNR, up to approximately 8 dB, than the open-loop system. These SNR improvements are measured consistently for all tested audio excerpts. Third, the SNR improvements achieved when using balanced-delay filterbank are also measured for 2, 4, and 5

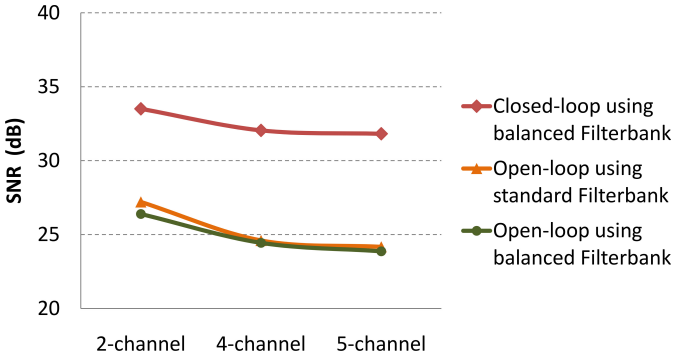

Fig. 4. Signal-to-Noise Ratio (SNR) measured on various filterbank for down-mixing 2-channel, 4-channel, and 5-channel audio signals

input channels. The results indicate that applying balanced-delay filterbank in the Closed-Loop MPS system potentially provides higher quality reconstructed audio signals.

VI. Conclusion

This paper has presented the delay characteristic of Quadrature Mirror Filterbank (QMF) that is specified in MPEG Surround (MPS) standard and then discussed a balanced-delay QMF that is very useful for application in a Closed-loop Spatial Audio Coding (CL-SAC). The experimental results have shown that employing the balanced-delay QMF in a CL-SAC can improve Signal-to-Noise Ratio (SNR) up to 8 dB in comparison to standard QMF (unbalanced-delay) in an open-loop MPS. This balanced-delay QMF provides an opportunity to apply MPS in a closed-loop configuration which is expected to be capable of minimising signal distortion during the processes of quantisation and encoding in MPS which then make it possible to operate MPS as bitrate scalable codec up to achieve perfect waveform reconstruction.

Acknowledgment

This work was funded by the Ministry of Education and Culture, the Republic of Indonesia under the scheme of Higher Education Excellent Research (Penelitian Unggulan Perguruan Tinggi). Furthermore, the authors thank the reviewers for their constructive comments and suggestions. The authors also say thank you to their colleagues in the department: Dr. Rahmadi Kurnia and Mrs. Fitrilina who also currently working in the project for helping in some helpful discussion during conducting the experiment and preparing the manuscript. We also thank the University of Surrey where earlier work on the Closed-Loop Spatial Audio Coding was carried out.

References

[1] K. Brandenburg, C. Faller, J. Herre, J. D. Johnston, and W. B. Kleijn, "Perceptual coding of high-quality digital audio," *Proceedings of the IEEE*, vol. 101 No. 9, pp. 1905–1919, 2014.

[2] T. Sugimoto, Y. Nakayama, and S. Oode, "Bitrate of 22.2 multichannel sound signal meeting broadcast quality," in *Proc. 137th AES Convention*, Los Angeles, USA, Oct. 2014.

[3] S. Kim, Y. Lee, and V. Pulkki, "New 10.2-channel vertical surround system (10.2-vss); comparison study of perceived audio quality in various multichannel sound systems with height loudspeakers," Presented at the 129^{th} Convention of the Audio Engineering Society, San Fransisco, USA, November 2010.

[4] A. Griffin, T. Hirvonen, C. Tzagkarakis, A. Mouchtaris, and P. Tsakalides, "Single-channel and multi-channel sinusoidal audio coding using compressed sensing," *IEEE Trans. on Audio, Speech, and Language Processing.*, vol. 19 no. 5, pp. 1382–1395, 2011.

[5] H. Moon, "A low-complexity design for an mp3 multichannel audio decoding system," *IEEE Trans. on Audio, Speech, and Lang. Proc.*, vol. 20, no. 1, pp. 314–321, January 2012.

[6] M. Nema and A. Malot, "Comparison of multichannel audio decoders for use in handheld devices," Presented at the 128^{th} Convention of the Audio Engineering Society, London, UK, May 2010.

[7] D. P. Chen, H. F. Hsiao, H. W. Hsu, and C. M. Liu, "Gram-schmidt-based downmixer and decorrelator in the MPEG surround coding," Presented at the 128^{th} Convention of the Audio Engineering Society, London, UK, May 2010.

[8] J. Hilpert and S. Disch, "The MPEG Surround audio coding standard [Standards in a nutshell]," *IEEE Signal Processing Mag.*, vol. 26, no. 1, pp. 148–152, Jan. 2009.

[9] J. Herre *et al.*, "MPEG Surround - The ISO/MPEG standard for efficient and compatible multichannel audio coding," *J. Audio Eng. Soc.*, vol. 56, no. 11, pp. 932–955, 2008.

[10] I. Elfitri, M. Muharam, and M. Shobirin, "Distortion analysis of hierarchical mixing technique on MPEG surround standard," in *Proc. of 2014 Int. Conf. on Advanced Computer Sciences and Information System*, Jakarta, Indonesia, October 2014.

[11] J. Herre, C. Faller, S. Disch, C. Ertel, J. Hilpert, A. Hoelzer, K. Linzmeier, C. Spenger, and P. Kroon, "Spatial audio coding: Next-generation efficient and compatible coding of multi-channel audio," in *Proc. the* 117^{th} *Convention of the Audio Engineering Society*, San Fransisco, CA, USA, Oct. 2004.

[12] J. Herre, "From joint stereo to spatial audio coding - recent progress and standardization," in *Proc. of the* 7^{th} *Int. Conf. on Digital Audio Effects (DAFx'04)*, Naples, Italy, October 2004.

[13] J. Herre and S. Disch, "New concepts in parametric coding of spatial audio: From SAC to SAOC," in *Proc. IEEE Int. Conf. on Multimedia and Expo*, San Fransisco, CA, USA, Oct. 2007.

[14] I. Elfitri, R. Kurnia, and D. Harneldi, "Experimental study on improved parametric stereo for bit rate scalable audio coding," in *Proc. of 2014 Int. Conf. on Information Tech. and Electrical Eng.*, Jogjakarta, Indonesia, October 2014.

[15] S. Gorlow, E. A. P. Habets, and S. Marchand, "Multichannel object-based audio coding with controllable quality," in *Proc. 2013 IEEE Int. Conf. Acoustics, Speech and Signal Proc.*, Vancouver, Canada, June 2013.

[16] J. Herre, C. Falch, D. Mahne, G. del Galdo, M. Kallinger, and O. Thiergart, "Interactive teleconferencing combining spatial audio object coding and DirAC technology," Presented at the 128^{th} Convention of the Audio Engineering Society, London, UK, May 2010.

[17] B. Günel, E. Ekmekçioğlu, and A. M. Kondoz, "Spatial synchronization of audiovisual objects by 3D audio object coding," in *Proc. of the IEEE International Workshop on Multimedia Signal Processing (MMSP)*, St. Malo, France, October 2010.

[18] J. Engdegard *et al.*, "Spatial audio object coding (SAOC)-The upcoming MPEG standard on parametric object based audio coding," Presented at the 124^{th} Convention of the Audio Engineering Society, Amsterdam, The Netherlands, May 2008.

[19] J. Herre, J. Hilpert, A. Kuntz, and J. Plogsties, "MPEG-H Audio The new standard for universal spatial/3d audio coding," *J. Audio Eng. Soc.*, vol. 62, no. 12, pp. 821–830, 2015.

[20] I. Elfitri, B. Gunel, and A. M. Kondoz, "Multichannel audio coding based on analysis by synthesis," *Proc. of the IEEE*, vol. 99, no. 4, pp. 657–670, April 2011.

[21] I. Elfitri, X. Shi, and A. M. Kondoz, "Analysis by synthesis spatial audio coding," *IET Signal Processing*, vol. 8, no. 1, pp. 30–38, February 2014.

[22] I. Elfitri, R. Kurnia, and Fitrilina, "Investigation on objective performance of closed-loop spatial audio coding," in *Proc. of 2014 Int. Conf. on Information Tech. and Electrical Eng.*, Jogjakarta, Indonesia, October 2014.

[23] J. Roden, J. Breebart, J. Hilpert, H. Purnhagen, E. Schuijers, J. Koppens, K. Linzmeier, and A. Holzer, "A study of the MPEG Surround quality versus bit-rate curve," Presented at the 123^{th} Convention of the Audio Engineering Society, New York, USA, Oct. 2007.

[24] J. Breebaart, G. Hotho, J. Koppens, E. Schuijers, W. Oomen, and S. V. de Par, "Background, concepts, and architecture for the recent MPEG Surround standard on multichannel audio compression," *J. Audio Eng. Soc.*, vol. 55, pp. 331–351, 2007.

[25] ISO/IEC, "Information Technology - Coding of audio-visual objects, Part 3: Audio," ISO/IEC 14496-3:2009(E), International Standards Organization, Geneva, Switzerland, 2009.

[26] ——, "Information Technology - MPEG Audio Technologies, Part 1: MPEG Surround," ISO/IEC 23003-1:2007(E), International Standards Organization, Geneva, Switzerland, 2007.